Journal of Memory and Language 62 (2010) 227–239

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Retrieval mode distinguishes the testing effect from the generation effect

Jeffrey D. Karpicke [a,*], Franklin M. Zaromb [b]

[a] Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907-2081, United States
[b] Department of Psychology, Washington University in St. Louis, St. Louis, MO 63130, United States

ARTICLE INFO

Article history:
Received 1 September 2009
revision received 11 November 2009
Available online 24 December 2009

Keywords:
Retrieval practice
Testing effect
Generation effect
Retrieval processes
Learning

ABSTRACT

A series of four experiments examined the effects of generation vs. retrieval practice on subsequent retention. Subjects were first exposed to a list of target words. Then the subjects were shown the targets again intact for Read trials or they were shown fragments of the targets. Subjects in Generate conditions were told to complete the fragments with the first word that came to mind while subjects in Recall conditions were told to use the fragments as retrieval cues to recall words that occurred in the first part of the experiment. The instruction manipulated retrieval mode—the Recall condition involved intentional retrieval while the Generate condition involved incidental retrieval. On a subsequent test of free recall or recognition, initial recall produced better retention than initial generation. Both generation and retrieval practice disrupted retention of order information, but retrieval enhanced retention of item-specific information to a greater extent than generation. There is a distinction between the testing effect and the generation effect and the distinction originates from retrieval mode. Intentional retrieval produces greater subsequent retention than generating targets under incidental retrieval instructions.

Introduction

A venerable way to study the nature of retrieval processes is to examine the effect of one retrieval on another. Experiments in which subjects repeatedly retrieve and reconstruct the past have a long history in memory research (e.g., Ballard, 1913; Bartlett, 1932; Brown, 1923; Gates, 1917; Tulving, 1967; see too Payne, 1987). Recently the effects of repeatedly testing memory have captured the attention of contemporary researchers who are interested in educational applications of retrieval practice (for overviews see McDaniel, Roediger, & McDermott, 2007; Metcalfe & Kornell, 2007; Pashler, Rohrer, Cepeda, & Carpenter, 2007; Roediger & Karpicke, 2006a). This renewed interest has also led to new examinations of the nature of the mnemonic effects of retrieval, which is the focus of this paper.

On the surface the testing effect seems to share many similarities with the generation effect (Jacoby, 1978; Slamecka & Graf, 1978). Consider the typical design of experiments that examine either the effects of retrieval practice (testing) or generation on subsequent retention. In most retrieval practice experiments, subjects study materials like word lists or text passages and take an initial test. The initial test often involves recall but recognition and multiple-choice tests have also been used. The effect of initial retrieval is assessed on a later criterial test which may be the same or different format as the first test and may occur relatively immediately or at a longer retention interval. The key finding is that practicing retrieval on the initial test enhances performance on the criterial test relative to a control condition where subjects repeatedly study to equate nominal exposure time to the materials in the two conditions (for review see Roediger & Karpicke, 2006a).

Now consider the design of a prototypical generation effect experiment. Here subjects are induced to generate items during the initial learning phase. This may be

* Corresponding author. Fax: +1 765 496 1264.
E-mail address: karpicke@purdue.edu (J.D. Karpicke).

0749-596X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2009.11.010

accomplished in a variety of ways—subjects might be asked to complete a fragment of a target word (*fr_ _nd*) or produce the target when given an antonym as a cue (generate *friend* when given *enemy* as a cue) or told to unscramble an anagram to form the target word (generate *friend* when given *fndrie*). The effect of generation is typically assessed on a criterial test of free recall or recognition. The key finding is that generation often enhances performance on the criterial test relative to a control condition where subjects read words intact (Jacoby, 1978; Slamecka & Graf, 1978; for review see Bertsch, Pesta, Wiscott, & McDaniel, 2007). There are important boundary conditions to the generation effect and we will discuss them shortly.

The question we asked in this research was this: Are there any meaningful differences between the testing effect and the generation effect? Many authors continue to lump together the testing effect and the generation effect with good reason—there is currently no well-developed empirical or theoretical basis to distinguish the effects (cf. Carrier & Pashler, 1992). In addition a number of researchers have espoused the benefits of retrieval practice for student learning (Karpicke & Roediger, 2008; McDaniel et al., 2007; Metcalfe & Kornell, 2007; Pashler et al., 2007) but this has not yet garnered wide support in education. In contrast the effectiveness of generative learning activities has been largely embraced in educational circles (see Mayer, 2008; see too Chi, 2000; King, 1994; Wittrock, 1974, 1989). We point this out because any distinction between generation and retrieval practice would have not only theoretical implications but also practical implications for learning in educational contexts.

Two observations motivated this research. The first observation is that the instructions given to subjects in generation effect experiments differ from the instructions in retrieval practice experiments. In most generation effect experiments the subjects are instructed to generate target items and can rely on any strategy that might accomplish this task. Many of the tasks used to induce generation (like completing a word fragment, or producing a word that is conceptually related to a cue, or unscrambling an anagram) are similar to implicit memory tests because these generation tasks do not involve intentional retrieval. Subjects are instructed to produce target items but are not required to think back to a prior episode or experience (Graf & Schacter, 1985; Roediger & McDermott, 1993; Schacter, 1987). In contrast, in experiments that examine retrieval practice the subjects are instructed to retrieve items that occurred in a study episode. Most retrieval practice tasks involve intentional retrieval—subjects are instructed to reconstruct knowledge about a study event that occurred in a particular place at a particular time (retrieve the words from a particular list or recall ideas from a particular text). Recall tests likely involve generation but generation does not necessarily involve recovering the spatiotemporal context in which an event occurred.

The difference between generation and retrieval instructions is essentially parallel to the distinction between incidental and intentional retrieval and constitutes a difference in what Tulving (1983) called *retrieval mode*. Subjects are thought to be in an episodic retrieval mode when told to think back to the past as they are on explicit memory tests. Moreover, subjects are thought to process retrieval cues differently in this cognitive state than they do under incidental retrieval conditions where subjects do not consciously think back to the past (as on implicit memory tests; see Graf & Schacter, 1985; Roediger & Blaxton, 1987). It is possible to hold all conditions and test cues constant and manipulate only retrieval mode by varying the instructions given to subjects (cf. to the retrieval intentionality criterion; see Schacter, Bowers, & Booker, 1989; see too Roediger, Weldon, Stadler, & Riegler, 1992). In the experiments reported here we examined whether the retrieval mode engaged in by subjects—intentional vs. incidental retrieval—would differentiate the testing effect from the generation effect.

The second observation that motivated this research is that generation effects are sensitive to aspects of experimental design that do not impact the testing effect. Specifically when the final criterial test involves free recall, generation effects are often found in mixed-list (within-subject) designs but not in pure-list (often between-subject) designs (see Begg & Snider, 1987; Hirshman & Bjork, 1988; Schmidt & Cherry, 1989; Slamecka & Katsaiti, 1987). On the contrary, testing effects are found in both within- and between-subject experiments when the final test involves free recall. For example, Carpenter, Pashler, and Vul (2006) and Roediger and Karpicke (2006b) obtained testing effects with both within- and between-subjects designs and with free recall as the criterial measure.

With respect to the generation effect, one explanation of the moderating influence of list composition is the item-order account first proposed by Nairne and colleagues (Nairne, Riegler, & Serra, 1991; see McDaniel & Bugg, 2008). This account is conceptually similar to other tradeoff or multifactor accounts of the generation effect, though differences between the accounts have been discussed elsewhere (see Hirshman & Bjork, 1988; McDaniel, Riegler, & Waddill, 1990; McDaniel, Waddill, & Einstein, 1988; Mulligan & Lozito, 2004). The idea behind the item-order account is that subjects encode attributes or features pertaining to the individual items in a list and to the order in which the items occurred. On a free recall test subjects use order information as a structure to guide retrieval of target candidates and use item information to discriminate which items actually occurred in a prior study episode (for elaboration of these ideas see Crowder, 1979; Hunt & Einstein, 1981; Hunt & McDaniel, 1993; Mandler, 1969; Nairne, 2006; Postman, 1972; Underwood, 1969). When subjects are required to generate items during learning this enhances the processing of item-specific features but disrupts the encoding of order information (Nairne et al., 1991). Therefore in mixed lists that contain both generated and read (intact) items, the generated items benefit from enhanced item processing and both types of item suffer from disrupted order processing. The result is a generation effect—better free recall of generated items than read items in mixed-list designs (Serra & Nairne, 1993; see too Gardiner & Arthurs, 1982; Slamecka & Graf, 1978; Slamecka & Katsaiti, 1987).

But the story is different in pure-list designs. A pure list of generated items benefits from enhanced item processing but suffers from disrupted order processing. In contrast, a pure list of read items does not benefit from enhanced item processing but also does not suffer from disrupted order processing. Thus there is often no difference in free recall of read vs. generated lists, and in fact sometimes there is an advantage of reading over generating (e.g. Nairne et al., 1991; Schmidt & Cherry, 1989). The enhanced item processing that occurs in a pure list of generated items is not sufficient to counteract disrupted order processing and produce an advantage in free recall relative to a pure list of read items.

Generation effects are clearly sensitive to experimental design but retrieval practice effects do not appear to depend on this factor. Several prior studies have shown that practicing retrieval produces greater retention than repeated study in pure-list, between-subject designs that employ final free recall (see Carpenter, 2009; Carpenter & DeLosh, 2006; Hogan & Kintsch, 1971; Karpicke & Roediger, 2007; Roediger & Karpicke, 2006b; Thompson, Wenger, & Bartling, 1978; Wheeler, Ewers, & Buonanno, 2003). The effects of list composition have not been examined as rigorously in the testing effect literature as they have been in the generation effect literature but a few studies have addressed the issue directly. Namely, Carpenter et al. (2006) examined both pure- and mixed-lists and showed positive testing effects on final free recall with both types of design (see too Carpenter, 2009).

In sum, we have two reasons to suspect there may be important differences between engaging in generation and practicing retrieval. First, generation conditions often involve incidental retrieval while retrieval practice conditions involve intentional retrieval. When subjects practice retrieval they must think back to and attempt to reconstruct what happened in a prior study episode. In contrast, subjects do not need to be in an episodic retrieval mode to complete a generation task. Second, generation effects depend on aspects of the experimental design in ways that retrieval practice effects do not. This is especially true when the criterial measure involves free recall. Adopting the perspective of the item-order framework might make it possible to identify the locus of any differences between generating and retrieving.

In the four experiments reported here we sought to determine whether manipulating retrieval mode—by giving subjects either incidental or intentional retrieval instructions—would distinguish the testing effect from the generation effect. Our aim was to hold all aspects of the procedure constant and manipulate only whether subjects incidentally generated or intentionally recalled during the critical generate/recall phase. This presented a handful of methodological challenges. First, subjects typically do not study items prior to generating them in most generation effect experiments but subjects do study items prior to recalling them on an initial test in testing effect experiments. Of course, including a study episode for a recall condition but not for a generate (or read) condition would confound the experiment. Thus the subjects in all conditions experienced the target words under incidental learning conditions in an initial exposure phase prior to the read/generate/recall manipulation.

Second, it was critical to create conditions where manipulating incidental vs. intentional retrieval would not affect performance on the initial test. Any difference in performance on the initial test would cloud interpretation of differences observed on a subsequent criterial test (for elaboration see Underwood's (1964) classic paper). Fortunately, several prior studies have demonstrated that it is possible to hold all test cues constant and manipulate only incidental vs. intentional retrieval instructions and observe virtually identical levels of performance in the two instructional conditions (e.g. see Geraci & Rajaram, 2002; Hamilton & Rajaram, 2001; Roediger et al., 1992). The materials and tasks used in the present experiments were designed to produce equivalent levels of performance in the initial "Generate" and "Recall" conditions (that is, under incidental or intentional retrieval instructions).

Finally, we suspected that using materials that afforded easy generation of target words (e.g. pairs of antonyms) would encourage subjects to use an incidental retrieval strategy rather than intentional retrieval even when subjects were instructed to do the latter. Thus the materials used in the present experiments were somewhat more difficult than materials commonly used in generation effect experiments. While many generation effect experiments see initial generation performance above 90%, initial performance was closer to 75% in the present experiments. The intent was to insure that subjects could successfully generate targets under incidental retrieval instructions but that subjects would in fact think back to the prior study episode when given intentional retrieval instructions.

The general procedure was similar in each experiment. In an initial exposure phase (Phase 1) subjects viewed a list of words (e.g., *love*, *diet*) under incidental learning conditions. Then in Phase 2 one of three things happened. In a Read condition the subjects read the intact target words paired with related cue words (e.g., *heart – love*, *eat – diet*). In Generate and Recall conditions the subjects were given fragments of the target words paired with cue words (e.g., *heart – l_v_*, *eat – di_ _*) and instructed to complete the fragment. The only difference between the Generate and Recall conditions was the instructions. Subjects in the Generate condition were told to complete the fragment with the first word that came to mind that successfully completed it. Subjects in the Recall condition, in contrast, were told to use the fragment as a cue to help them recall a word that occurred in the first part of the experiment. Thus subjects in the Recall condition were placed in an episodic retrieval mode while subjects in the Generate condition were not. Finally, in Phase 3 the subjects were given a criterial test of free recall (Experiments 1 and 2) or recognition (Experiments 3 and 4).

Experiment 1

The purpose of Experiment 1 was to see if practicing retrieval would produce effects on future retention that differed from the effects of generation. Subjects either read, or generated, or recalled word pairs. A pure-list between-subjects design was used and free recall was the criterial measure. As described above, there are often no generation

effects in such designs but testing effects are observed in similar designs. In Experiment 1 the cues on the initial test were held constant. The only difference between the Generate and Recall conditions was whether subjects were given incidental or intentional retrieval instructions. If retrieval modes produce different effects on subsequent free recall, then there should be an advantage of engaging in intentional retrieval in the Recall condition relative to incidental retrieval in the Generate condition.

Method

Subjects

Sixty Purdue University undergraduates participated in Experiment 1 in exchange for course credit.

Materials

Forty word pairs were selected from Jacoby's (1996) norms. Each pair included a cue, a fragment, and two possible target words that completed the fragment (e.g., *heart – l_v_* could be completed with *love* or *live*; *eat – di_ _* could be completed with *diet* or *dine*). The pairs used in the present experiments had one target with a high completion baserate based on Jacoby's norms ($M = .65$) and an alternative with a low completion baserate ($M = .16$). In the present example *love* and *diet* are high baserate targets and *live* and *dine* are the low baserate alternatives. Only the high baserate targets were presented to subjects. Importantly, only the high baserate targets were counted as "correct" in the analyses in all subsequent experiments. Separate analyses of the alternate completions are also reported in each experiment. We chose these stimuli for two reasons. First, preliminary pilot work showed that subjects were able to come up with targets that completed these word fragments fairly frequently and easily. Second, the pilot work showed that completion rates were roughly equivalent in Generate and Recall conditions (that is, under incidental and intentional retrieval conditions; cf. Roediger et al., 1992).

Design

Experiment 1 used a pure-list between-subjects design. There were three conditions—Read, Generate, and Recall—and 20 subjects were assigned to each condition.

Procedure

The experiment comprised three critical phases and two filler tasks. In Phase 1 subjects saw a list of 40 target words (e.g., *love*, *diet*) presented on the computer screen one at a time for 2 s each with a 500 ms interstimulus interval. The subjects were given incidental learning instructions and told to read each word silently. They were not instructed to try to remember the words. Phase 1 was followed by a brief distracter period in which subjects completed the short form of the Need for Cognition scale (Cacioppo, Petty, & Kao, 1984). This questionnaire includes 18 questions designed to assess a person's tendency to engage in effortful cognitive activities. Performance on this task was not relevant to the current experiment. The distracter period lasted about 5 min.

Phase 2 was the critical phase where the Read/Generate/Recall manipulation occurred. Subjects in the Read condition studied the target words they saw in Phase 1 paired with an associated cue word (e.g., *heart – love*, *eat – diet*). Each word pair was shown for 4 s. The timing was chosen based on preliminary pilot testing that showed that this was about the same amount of time subjects needed to produce targets in the Generate and Recall conditions. Again the subjects were given incidental learning instructions and were simply told to read each pair silently. In the Generate and Recall conditions the subjects were shown the cue words and fragments of the target words (e.g., *heart – l_v_*, *eat – di_ _*). The Generate and Recall conditions differed only in the instructions given to subjects. Subjects in the Generate condition were told to type the first word that came to mind that was related to the intact cue word and successfully completed the fragment. Subjects in the Recall condition were told that Phase 2 was a recall test. They were instructed to use the intact cue word and fragment as clues to help them recall a word from Phase 1. Thus the subjects in the Recall condition were placed in retrieval mode—they were told to think back to the study phase and try to recall a word that completed the fragment (Roediger & Blaxton, 1987; Schacter et al., 1989; Tulving, 1983).

The generate/recall trials were self-paced. Subjects were instructed to press Enter to advance to the next trial after they produced a response. Response times were recorded as the total trial time—the time between the onset of the generate/recall trial and the Enter keypress. The subjects were encouraged to try hard to come up with a completion for each fragment but because the task was self-paced they were also told not to spend enormous amounts of time trying to come up with completions. The subjects could press Enter to move on to the next trial if they felt they could not come up with a completion. The computer was set to move on to the next trial automatically after 20 s.

Phase 2 was followed by another brief distracter task in which the subjects answered a series of 20 fictional general knowledge questions. Once again performance on this task was not relevant to the current experiment. The distracter period lasted about 5 min.

In Phase 3 subjects took a final free recall test. They were told to try to recall as many target words as they could, and they were told that the targets were the words they saw in the first part of the experiment and the words they read/generated/recalled in the second part. Subjects typed their responses on the computer and pressed Enter after typing each response. Their responses remained displayed in a list on the screen throughout the recall period (cf. Karpicke & Roediger, 2007). The free recall test lasted 7 min. At the end of the experiment the subjects were debriefed and thanked for their participation.

Results

Initial generation/recall

Table 1 shows the proportion of targets correctly produced in the initial generate/recall phase. The proportion

Table 1
Results of Experiment 1: Proportion of targets produced in initial generation/recall phase and proportion of targets recalled on final free recall test.

Condition	Proportion produced	Final recall
Read	–	.27 (.03)
Generate	.71 (.03)	.28 (.02)
Recall	.72 (.03)	.38 (.02)

Note: Standard errors are in parentheses.

was nearly identical in the Generate and Recall conditions (.71 vs. .72, $F < 1$). Response times for correct responses (correctly generated or recalled targets) averaged 4.12 s and 4.04 s in the Generate and Recall conditions respectively ($F < 1$). Neither mean was significantly different from the 4 s presentation rate in the Read condition (Fs < 1). Finally the proportion of alternate targets produced (the low baserate completions which were never seen in the experiment) was .11 and there was not a significant difference in the proportion produced in the Generate and Recall conditions (.12 vs. .09, $F < 1$).

Final free recall

The critical result of Experiment 1 is performance on the final free recall test shown in the right column of Table 1. These data reflect the proportion of targets recalled on the final test—thus the same set of items (the targets, not the alternatives) is being examined in each condition. There was not a significant advantage of generating vs. reading ($F < 1$). This replicates prior research using pure-list designs and final recall tests (e.g. Nairne et al., 1991; Schmidt & Cherry, 1989). However, engaging in intentional retrieval in the Recall condition produced greater final recall relative to reading ($F(1, 38) = 9.56$, $\eta_p^2 = .20$). This finding replicates prior work examining retrieval practice with pure-list between-subjects designs (Carpenter, 2009; Carpenter et al., 2006; Karpicke & Roediger, 2007; Roediger & Karpicke, 2006b). Finally and importantly, retrieval practice produced greater final recall relative to generating ($F(1, 38) = 8.30$, $\eta_p^2 = .18$). Thus retrieval mode—whether subjects were given intentional or incidental retrieval instructions—distinguished the effects of retrieval practice from the effects of generating target items.

Table 2 shows the results of an analysis of the relationship between initial generation/recall and final free recall. Following Tulving's (1964) convention for examining the fate of individual items across two tests, C_1 refers to items generated or recalled in the initial generate/recall phase and N_1 refers to items that were not generated or recalled in the initial phase. C_2 refers to items recalled on the final free recall test and N_2 refers to items not recalled on the final test. The top portion of Table 2 shows the joint probabilities for targets correctly produced in the initial generate/recall phase. There was greater intertest retention (C_1C_2) and less intertest forgetting (C_1N_2) in the Recall condition vs. the Generate condition (for C_1C_2, $F(1, 38) = 7.40$, $\eta_p^2 = .16$, and for C_1N_2, $F(1, 38) = 4.31$, $\eta_p^2 = .10$). The other means (N_1C_2 and N_1N_2) did not differ across conditions ($F < 1$). The bottom portion of Table 2 shows the joint probabilities for alternative completions

Table 2
Fates of individual items in Experiment 1: Joint probabilities between initial generation/recall and final free recall.

	C_1C_2	C_1N_2	N_1C_2	N_1N_2
Target completions (correct)				
Generate	.25 (.03)	.46 (.02)	.03 (.01)	.26 (.03)
Recall	.35 (.02)	.39 (.03)	.03 (.01)	.24 (.03)
Alternate completions (incorrect)				
Generate	.02 (.01)	.10 (.01)	.01 (.00)	.87 (.01)
Recall	.02 (.00)	.08 (.01)	.01 (.00)	.90 (.01)

Note: Standard errors are in parentheses. C_1 = items successfully generated or recalled in the initial generate/recall phase. N_1 = items that were not generated or recalled in the initial phase. C_2 = items successfully recalled on the final free recall test. N_2 = items not recalled on the final free recall test.

produced in the initial generate/recall phase. This analysis makes it possible to examine whether the production of alternatives interfered with final recall and if any interference occurred disproportionately in the Generate or Recall condition. As mentioned earlier the overall proportion of alternatives produced was 11%. Very few alternatives were produced and then recalled on the final test (C_1C_1 averaged about 2% in each condition) and there was no difference between Generate and Recall conditions ($F < 1$). There was a small (2%) difference in intertest forgetting of alternatives (C_1N_2, $F(1, 38) = 3.29$, $\eta_p^2 = .08$, $p = .08$). N_1C_2 did not differ across conditions ($F < 1$) and there was a small (3%) difference in N_1N_2 ($F(1, 38) = 2.81$, $\eta_p^2 = .07$, $p = .10$).

Discussion

The key finding in Experiment 1 was that there was no generation effect but there was a testing effect. That is, in a pure-list between-subject design, practicing retrieval produced a significant advantage in final free recall relative to reading but generating produced no effect. These patterns of results were suggested by prior research and they are captured here within a single experiment. The independent variable that distinguished retrieval practice from generation was retrieval mode—whether subjects were given intentional or incidental retrieval instructions. It is important to note that the effects on final recall cannot be attributed to differences in initial performance. The proportion of items correctly completed and the amount of time required to complete items did not differ across the two instructional conditions. Thus holding all test cues constant and manipulating only retrieval mode did not affect initial performance but produced different effects on subsequent free recall.

Experiment 2

Experiment 1 showed that there is a clear difference between retrieval practice and generation and that the difference depends on intentional vs. incidental retrieval instructions. Indeed Experiment 1 demonstrated a scenario where a "generative" learning task produced no benefit over reading but retrieval practice produced a significant

benefit. At this point we do not know exactly what is responsible for the different mnemonic effects of generating vs. retrieving. In the context of the item-order framework, the advantage of retrieval practice could be produced by enhanced order memory or enhanced item memory—both of which are necessary for free recall. That is, retrieval practice could enhance the retention of information about order required to develop a retrieval structure to support free recall. Alternatively retrieval practice could produce an even greater enhancement in item-specific processing than generation—enough of an enhancement to overcome a disruption in order processing (if order is indeed disrupted by retrieval practice). And of course the advantage of retrieval practice could arise because of a combination of enhanced item and order retention. The subsequent experiments were carried out to provide further evidence about the locus of the difference between retrieval practice and generation.

Experiment 2 was designed to replicate Experiment 1 with a mixed-list design. In Phase 2 half the items were presented intact in Read trials and half were presented as fragments in Generate or Recall trials. The instruction to generate or recall items was manipulated between-subjects. As in Experiment 1 the criterial test involved free recall. As noted earlier, the generation effect is typically observed in final free recall using mixed-list within-subjects designs (e.g. Serra & Nairne, 1993; Slamecka & Katsaiti, 1987). This occurs because generation enhances individual item processing while processing of order information is equally disrupted for read and generated items in a mixed list. Therefore we expected to find a generation effect in Experiment 2.

The critical question in Experiment 2 was whether intentional retrieval in the Recall condition would alter the pattern of results. One possibility is that generating and recalling enhance item-specific processing equivalently but recalling also enhances retention of order information. If true then there might be no advantage of the Recall condition relative to the Generate condition in a mixed-list design. However, enhanced order processing in the Recall condition might also occur for the Read items studied by those subjects. This might produce a difference in final recall of Read items across the two instructional conditions. On the other hand, the advantage of retrieval practice might stem from enhanced item processing that is even greater than the enhancement from generation alone (great enough to overcome a disruption of order). In that case we would expect to see an advantage of Recall over Generate but perhaps no difference in recall of Read items across instructional conditions.

Method

Subjects

Forty Purdue University undergraduates participated in Experiment 2 in exchange for course credit. None had participated in Experiment 1.

Materials

The 40 items used in Experiment 1 were used again in Experiment 2. The materials were divided into two sets of 20 items. The two sets were equated in terms of the baserates for completing the fragment with the targets (.66 and .65 respectively) and completing with alternatives (.17 and .15 respectively). The assignment of sets to conditions was counterbalanced across subjects.

Design

Experiment 2 used a mixed-list design. In Phase 2 half the items were presented intact for Read trials and half were presented as fragments for Generate/Recall trials. Thus Read vs. Generate/Recall was manipulated within-subjects. Generate vs. Recall was manipulated between-subjects. Twenty subjects were given Generate instructions and 20 were given Recall instructions. The instructions given to subjects were identical to those used in Experiment 1.

Procedure

The procedure was identical to the procedure in Experiment 1 with one exception. In Phase 2 subjects were told that for half the pairs they would see the target word intact and for the other half they would see a fragment of the target. Subjects in the Generate group were told to complete the fragment with the first word that came to mind. Subjects in the Recall group were told to use the cue and target as clues to help them recall a word from the previous exposure phase (Phase 1).

Results

Initial generation/recall

Table 3 shows the proportion of targets correctly produced in the initial generate/recall phase. As was true in Experiment 2 the proportion was nearly identical in the Generate and Recall conditions (.76 vs. .75, $F < 1$). Response times for correct responses averaged 4.68 s and 4.69 s in the Generate and Recall conditions respectively ($F < 1$). The mean response time in Experiment 2 (4.68 s) was significantly greater than the 4 s presentation rate in the Read condition ($F(1, 39) = 7.34$, $\eta_p^2 = .16$). The proportion of alternate targets produced was .13 and did not differ in the Generate vs. Recall conditions (.13 vs. .13, $F < 1$).

Final free recall

The key data in Experiment 2 are the final free recall data shown in the right column of Table 3. There was an advantage of generating vs. reading items and an even

Table 3
Results of Experiment 2: Proportion of targets produced in initial generation/recall phase and proportion of targets recalled on final free recall test.

Condition	Proportion produced	Final recall
Generate group		
Read	–	.22 (.02)
Generate	.76 (.02)	.32 (.03)
Recall group		
Read	–	.22 (.03)
Recall	.75 (.03)	.40 (.03)

Note: Standard errors are in parentheses.

greater advantage of recalling vs. reading items. The data were entered into a 2 (Item Type: Read vs. Generate/Recall) × 2 (Instruction: Generate vs. Recall) ANOVA. There was a main effect of item type ($F(1, 38) = 39.31$, $\eta_p^2 = .51$). The main effect of condition did not reach significance ($F(1, 38) = 2.36$, $\eta_p^2 = .06$, $p = .13$) but there was a marginally significant interaction ($F(1, 38) = 3.48$, $\eta_p^2 = .08$, $p = .07$). Pairwise comparisons showed that there was no difference in the Read items in the two groups ($F < 1$). There was a significant generation effect in the Generate group ($F(1, 19) = 12.15$, $\eta_p^2 = .39$). Out of 20 subjects in the Generate group, 13 showed Generate > Read, 4 showed Read > Generate and there were 3 ties. There was also a significant testing effect in the Recall group ($F(1, 19) = 27.54$, $\eta_p^2 = .59$). Out of 20 subjects in the Recall group, 16 showed Recall > Read, 2 showed Read > Recall and there were 2 ties. Finally and importantly, final free recall was greater in the Recall condition than in the Generate condition ($F(1, 38) = 5.45$, $\eta_p^2 = .13$) replicating the advantage of intentional retrieval seen in Experiment 1.

Table 4 shows the analysis of the fate of individual items across two tests. The pattern conceptually replicates Experiment 1. Intertest retention (C_1C_2) was 6% greater in the Recall condition than in the Generate condition but this difference did not reach significance ($F(1, 38) = 2.20$, $\eta_p^2 = .06$, $p = .15$). Intertest forgetting (C_1N_2) was 7% greater in the Generate condition relative to the Recall condition and this difference approached significance ($F(1, 38) = 3.22$, $\eta_p^2 = .08$, $p = .08$). There were no differences in N_1C_2 or N_1N_2 (Fs < 1). As in Experiment 1, even though some alternate completions were produced initially they were not frequently recalled on the final test (for alternate completions C_1C_2 averaged 2%). For the alternate completion data all Fs were less than 1.

Discussion

Experiment 2 provides a conceptual replication of Experiment 1 with a mixed-list design. There was a generation effect, replicating prior work with mixed lists, and there was also a retrieval practice effect. But most importantly retrieval practice produced greater final recall than generating. Based on the item-order tradeoff theory, the generation effect occurs in a mixed list because generation enhances item-specific processing but disrupts retention of order information for both generated and read items. If retrieval practice enhanced retention of order information we might expect this to occur for all items in the list—including read items—and thereby produce a difference in retention of read items in the Generate and Recall conditions. But this did not occur. Instead the data are more consistent with the idea that retrieval practice enhanced item-specific processing to an even greater extent than generation.

Table 4
Fates of individual items in Experiment 2: Joint probabilities between initial generation/recall and final free recall.

	C_1C_2	C_1N_2	N_1C_2	N_1N_2
Target completions (correct)				
Generate	.31 (.03)	.45 (.03)	.02 (.01)	.22 (.02)
Recall	.37 (.03)	.38 (.03)	.03 (.01)	.23 (.02)
Alternate completions (incorrect)				
Generate	.02 (.01)	.11 (.02)	.01 (.00)	.87 (.02)
Recall	.02 (.01)	.11 (.01)	.01 (.00)	.87 (.02)

Note: Standard errors are in parentheses. C_1 = items successfully generated or recalled in the initial generate/recall phase. N_1 = items that were not generated or recalled in the initial phase. C_2 = items successfully recalled on the final free recall test. N_2 = items not recalled on the final free recall test.

Experiment 3

Experiments 1 and 2 established that retrieval mode distinguishes the testing effect from the generation effect when the final criterial test involves free recall. The purpose of Experiment 3 was twofold. The first purpose was to see if the superiority of retrieval practice to generation would also be observed in a final item recognition memory test, a test which is presumably more sensitive to item information than to order information (Hunt & Einstein, 1981; Nairne et al., 1991). The second purpose was to examine the effects of generation and retrieval practice on order memory by using a reconstruction of order test (Nairne et al., 1991; see too Greene, Thapar, & Westerman, 1998; Mulligan, 2002; Serra & Nairne, 1993). On this test subjects are shown items and told to put them in the order in which they were presented. An order reconstruction test is presumably more sensitive to retention of order information than to item information. Of course the test is not a "pure" assessment of order memory in the absence of item memory because presenting subjects with the items provides them only with copy cues of the original studied items (Tulving & Thomson, 1973) and does not necessarily reinstate memory for the occurrence of the items.

The procedure in Experiment 3 was similar to the ones used in the previous experiments. Subjects were first exposed to the list of targets in Phase 1. In Phase 2 the subjects either read, or generated, or recalled the target items. A pure-list between-subject design was used (the same design as Experiment 1). In light of the response time results of Experiment 2 the presentation rate in the Read condition was increased to 4.5 s. In Phase 2 the list was presented in 8-item sets and after each set the subjects either completed a distracter task or completed an order reconstruction test. The items that were not tested on the order reconstruction tests in Phase 2 were tested on a final yes/no recognition test in Phase 3. This procedure was modeled after the one used by Nairne et al. (1991).

Prior studies have found generation effects on recognition tests with pure-list designs although it is worth noting that the sizes of the effects are sometimes small in such experiments (e.g. Begg & Snider, 1987; Nairne et al., 1991). Prior studies have also shown that performance on order reconstruction tests is greater for pure-lists of read items than for pure-lists of generated items (e.g. Greene et al., 1998; Mulligan, 2002; Nairne et al., 1991). We expected to replicate this pattern. The key question in Experiment 3 was how intentional retrieval in the Recall condition would impact the results. If retrieval practice enhances retention of order information then this enhancement should appear on the order reconstruction test and

performance might be better in the Recall condition relative to the Generate condition (and perhaps relative to the Read condition as well). Likewise if retrieval practice primarily enhances item-specific processing then the Recall condition should outperform the Generate condition on the final recognition test. Of course retrieval practice might enhance retention of both item and order information and show effects on both types of test.

Method

Subjects

Sixty Purdue University undergraduates participated in Experiment 3 in exchange for course credit. None of the subjects had participated in the prior experiments.

Materials

Forty-eight pairs were used in Experiment 3. Eight new pairs from Jacoby's (1996) norms were added to the set of 40 used in the previous experiments. For the set of 48 pairs the mean completion baserate for targets was .63 and the baserate for alternatives was .16.

Design

Experiment 3 used a pure-list design. Condition (Read, Generate, or Recall) was manipulated between-subjects and 20 subjects were assigned to each condition. Test format (Order Reconstruction vs. Recognition) was manipulated within-subjects but between-materials. The 48 pairs were divided into six sets of eight pairs. The six sets were equated in terms of the baserates for completing the fragment with the targets or alternatives. Three of the six sets were assigned to the order reconstruction test and the other three were assigned to the recognition test. The assignment of sets to test format was counterbalanced across subjects.

Procedure

The procedure was similar to the procedures used in the previous experiments. Subjects were first exposed to the list of 48 target words in Phase 1 and this phase was followed by the Need for Cognition questionnaire as a distracter task. In Phase 2 the subjects were shown the 48 cues and intact target words in the Read condition or cues and fragments of the target words in the Generate and Recall conditions. Items were presented for 4.5 s in the Read condition. As in the previous experiments the Generate and Recall groups differed only in the instructions given to the subjects. Subjects in the Generate group were told to complete the fragment with the first word that came to mind. Subjects in the Recall group were told to use the cue and target as clues to help them recall a word from the previous exposure phase (Phase 1).

The subjects were told that after every set of eight pairs they would perform a brief task on the computer that involved verifying multiplication problems. This distracter task lasted 30 s. Following three of the six sets the subjects did not take an order reconstruction test and the computer advanced to the next set. The items in these sets were tested on the final recognition test. Following the other three sets the subjects completed an order reconstruction test. They were shown the eight pairs they had just seen simultaneously in a column on the computer screen but in a new random order. The subjects were told to write down the items in their original order of presentation. They made their responses in a response book prepared for them which contained three sheets of paper with a column of eight lines on each sheet. The subjects were instructed to fill in each response line on the sheet and not to repeat items. The order reconstruction tests were self-paced. When subjects completed a test they pressed the F1 key to advance to the next set of 8 word pairs and turned to the next page in their response book.

Phase 2 was followed by the brief distracter task wherein subjects answered fictional general knowledge questions. Then in Phase 3 subjects took a final old/new recognition test. The test comprised 48 words: the 24 target words from the three sets that were not tested in Phase 2 and 24 new distracter words. The distracter words were obtained based on data from the English Lexicon Project (Balota et al., 2007) such that the 24 distracters were matched with the total set of 48 target words in terms of word length, word frequency, and orthographic neighborhood. On each trial in the recognition test the subjects were shown a word and asked to decide if they had seen the word earlier in the experiment. The subjects pressed 1 to indicate "old" and 0 to indicate "new". The test was self-paced and subjects were required to make a response to each test item. At the end of the experiment the subjects were debriefed and excused.

Results

Initial generation/recall

Table 5 shows the proportion of targets correctly produced in the initial generate/recall phase. The proportion was nearly identical in the Generate and Recall conditions (.73 vs. .75, $F < 1$). Response times for correct responses averaged 4.52 s and 4.69 s in the Generate and Recall conditions respectively ($F < 1$). The mean response time in Experiment 2 (4.61 s) was not significantly greater than the 4.5 s presentation rate in the Read condition ($F < 1$). The proportion of alternate targets produced was .12 and there was not a significant difference between the Generate and Recall conditions (.13 vs. .11, $F(1, 38) = 1.42$, n.s.).

Order reconstruction

Table 5 also shows the results of the order reconstruction task. The means reflect the proportion of items placed

Table 5
Results of Experiment 3: Proportion of targets produced in initial generation/recall phase and proportion of targets recognized (hits) and lures mistakenly identified (false alarms) on the final recognition test.

Condition	Proportion produced	Order reconstruction	Probability "old"	
			Targets (hits)	Lures (false alarms)
Read	–	.56 (.06)	.79 (.02)	.21 (.04)
Generate	.73 (.02)	.26 (.03)	.85 (.02)	.20 (.03)
Recall	.75 (.02)	.27 (.04)	.89 (.02)	.16 (.03)

Note: Standard errors are in parentheses.

in their correct serial position and the data were scored without regard for generation/recall success or failure. In a separate analysis the data were conditionalized on successful generation/recall and the means obtained for the Generate and Recall conditions were .28 and .26 respectively. Neither was significantly different from the nonconditionalized data (Fs < 1) and thus the subsequent analyses were performed on the nonconditionalized data.

The data in Table 5 clearly show that the Read condition outperformed both the Generate and Recall conditions on the order reconstruction test. Pairwise comparisons confirmed that order reconstruction performance was superior in the Read condition relative to the Generate condition ($F(1, 38) = 21.54$, $\eta_p^2 = .36$) and the Recall condition ($F(1, 38) = 16.08$, $\eta_p^2 = .30$). There was no difference in performance in the Generate and Recall conditions ($F < 1$). A final analysis of the order reconstruction data was carried out to examine performance across serial positions (cf. Nairne et al., 1991). Fig. 1 shows order performance as a function of serial position. These data were entered into a 3 (Condition) × 8 (Serial Position) ANOVA. There was a main effect of Condition ($F(2, 57) = 14.75$, $\eta_p^2 = .34$) and a main effect of Serial Position ($F(7, 399) = 18.34$, $\eta_p^2 = .24$) but no interaction ($F(14, 399) = 1.06$, n.s.). Thus the disruption in processing order information occurred equivalently across serial positions for the Generate and Recall conditions.

Final recognition

Table 5 shows the proportion of items called "old" on the final recognition test (hits and false alarms). A oneway ANOVA showed an overall effect of condition on hits ($F(2, 57) = 6.82$, $\eta_p^2 = .19$) and a separate oneway ANOVA showed no effect on false alarms ($F < 1$). Pairwise comparisons were performed on the hit rates. There was a generation effect—hit rates were greater in the Generate condition than in the Read condition ($F(1, 38) = 4.31$, $\eta_p^2 = .10$). There was also a testing effect—the Recall condition outperformed the Read condition ($F(1, 38) = 15.94$, $\eta_p^2 = .30$). Finally, there was a 4% difference between the hit rates in the Recall and Generate conditions but this difference did not reach significance ($F(1, 38) = 2.07$, $\eta_p^2 = .05$, $p = .16$, two-tailed). It is worth noting that when false alarms were subtracted from hits the advantage of Recall relative to Generate was marginally significant (.73 vs. .65, $F(1, 38) = 3.19$, $\eta_p^2 = .08$, $p = .08$, two-tailed).

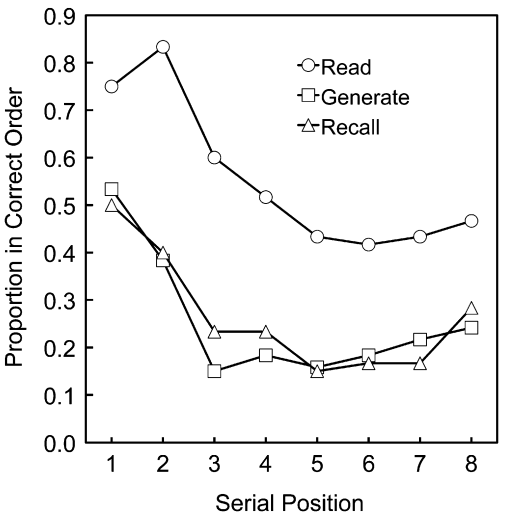

Fig. 1. Order reconstruction performance for the Read, Generate, and Recall conditions in Experiment 3 plotted as a function of serial position.

Table 6 shows the analysis of the fate of individual items across two tests (initial generation/recall and subsequent recognition). Intertest retention (C_1C_2) was 4% greater in Recall than Generate though this difference did not reach significance ($F(1, 38) = 1.18$, n.s.). There was not a significant difference in intertest forgetting (C_1N_2) across the two conditions ($F(1, 38) = 1.42$, n.s.) nor were there differences in N_1C_2 or N_1N_2 (Fs < 1). An important question (similar to one asked in the prior experiments) is whether generation of the alternate completions would interfere with recognition of targets disproportionately in the Generate condition relative to the Recall condition. The bottom portion of Table 6 shows the relevant data and shows that this was not the case. When an alternative was produced, correct recognition of the target (C_1C_2) and failure to recognize the target (C_1N_2) did not differ across conditions (both Fs < 1).

Discussion

Experiment 3 extended the findings from Experiments 1 and 2 to final recognition. Both generation and retrieval practice produced positive effects on subsequent recognition. There was also an advantage of retrieval practice relative to generation, conceptually replicating the results of Experiment 1 and 2, though the difference did not reach significance. Importantly, performance on the order reconstruction test was better in the Read condition than in the Generate or Recall conditions which did not themselves differ. This pattern of results lends further support to the idea that retrieval practice disrupts retention of order information, just like generation, but enhances item-specific processing to a greater extent than generation.

Experiment 4

The purpose of the final experiment was to examine the effect of reading, generating, or recalling items on final rec-

Table 6
Fates of individual items in Experiment 3: Joint probabilities between initial generation/recall and identifying targets as "old" on the final recognition test.

	C_1C_2	C_1N_2	N_1C_2	N_1N_2
Target completions (correct)				
Generate	.69 (.03)	.03 (.01)	.16 (.02)	.12 (.02)
Recall	.73 (.03)	.02 (.01)	.16 (.02)	.09 (.02)
Alternate completions (incorrect)				
Generate	.09 (.01)	.05 (.01)		
Recall	.08 (.01)	.04 (.01)		

Note: Standard errors are in parentheses. C_1 = items successfully generated or recalled in the initial generate/recall phase. N_1 = items that were not generated or recalled in the initial phase. C_2 = target items called "old" (hits) on the final recognition test. N_2 = target items called "new" (misses) on the final recognition test.

ognition using a mixed-list design. In Phase 2 half the items were presented intact in Read trials and half were presented as fragments in Generate or Recall trials. The instruction to generate or recall items was manipulated between-subjects (just as was done in Experiment 2). The criterial test in Phase 3 was a yes/no recognition test. The prediction was that there would be a significant generation effect on final recognition, replicating prior research that also used mixed-list designs. In addition, if intentional retrieval in the Recall condition enhances item-specific processing (as suggested by the results of Experiments 1–3) then final recognition should be better in the Recall condition than in the Generate condition.

Method

Subjects

40 Purdue University undergraduates participated in Experiment 4 in exchange for course credit. None of the subjects had participated in the prior experiments.

Materials

The 48 pairs used in Experiment 3 were used in Experiment 4.

Design

Experiment 4 used the same mixed-list design used in Experiment 2. In Phase 2 half the items were presented intact for Read trials and half were presented as fragments for Generate/Recall trials. Thus Read vs. Generate/Recall was manipulated within-subjects while Generate vs. Recall was manipulated between-subjects. Twenty subjects were given Generate instructions and 20 were given Recall instructions.

Procedure

The procedure was identical to the procedure used in Experiment 3 with three exceptions. First, in Phase 2 there were no order reconstruction tests. All 48 items were presented in Phase 2 in a random order. Second, in Phase 2 half the items were presented intact for Read trials and half were presented as fragments for Generate or Recall trials (as was done in Experiment 2). Third, the final old/new recognition test was identical to the one used in Experiment 3 but included 96 words: the 48 target words (24 Read and 24 Generate/Recall) and 48 distracter words. The distracters included the 24 distracters used in Experiment 3 plus 24 additional distracters that were also matched with the set of 48 target words in terms of word length, word frequency, and orthographic neighborhood based on data from the English Lexicon Project (Balota et al., 2007).

Results

Initial generation/recall

Table 7 shows the proportion of targets correctly produced in the initial generate/recall phase. There was no difference between the Generate and Recall conditions (.77 vs. .77, $F < 1$). Response times for correct responses averaged 4.47 s and 4.90 s in the Generate and Recall conditions respectively ($F < 1$). The mean response time in Experiment 2 (4.65 s) was not significantly greater than the 4.5 s presentation rate for Read items ($F < 1$). Finally, the proportion of alternate targets produced was .13 and did not differ in the Generate vs. Recall conditions (.13 vs. .12, $F < 1$).

Final recognition

Table 7 shows final recognition. First, there was no difference in false recognition of lures (.28 vs. .31, $F < 1$). The data from the Generate and Recall conditions were entered into a 2 (Item Type: Read vs. Generate/Recall) × 2 (Instruction: Generate vs. Recall) ANOVA. There was a main effect of item type ($F(1, 38) = 27.67$, $\eta_p^2 = .42$). The main effect of instruction condition did not reach significance ($F(1, 38) = 2.57$, $\eta_p^2 = .06$, $p = .11$) but the instruction × item-type interaction was marginally significant ($F(1, 38) = 3.05$, $\eta_p^2 = .07$, $p = .08$). Pairwise comparisons showed that there was no difference in recognition of Read items in the two groups (.74 vs. .76, $F < 1$). There was a significant generation effect in the Generate group (.81 vs. .74, $F(1, 19) = 5.92$, $\eta_p^2 = .24$). Out of 20 subjects in the Generate group, 14 showed Generate > Read, 3 showed Read > Generate and there were 3 ties. There was also a significant testing effect in the Recall group (.90 vs. .76, $F(1, 19) = 25.69$, $\eta_p^2 = .58$). Out of 20 subjects in the Recall group, 18 subjects showed Recall > Read, 1 subject showed Read > Recall and there was 1 tie. Most importantly, there was a significant advantage of Recall over Generate (.90 vs. .81, $F(1, 38) = 4.87$, $\eta_p^2 = .11$) replicating the key results of the previous three experiments.

Table 8 shows the analysis of the fate of individual items across two tests. The data are consistent with the patterns observed in the previous experiments. The Recall group showed somewhat greater intertest retention (C_1C_2) than the Generate group (.74 vs. .68, $F(1, 38) = 3.12$, $\eta_p^2 = .08$, $p = .08$) and less intertest forgetting (C_1N_2, .03 vs. .09, $F(1, 38) = 5.40$, $\eta_p^2 = .12$). There were no differences in N_1C_2 or N_1N_2 (for N_1C_2 $F < 1$, for N_1N_2 $F(1, 38) = 1.38$, n.s.). Finally, when an alternate completion was produced initially, correct recognition of the target (C_1C_2) and failure to recognize the target (C_1N_2) did not differ across conditions (both Fs < 1).

Table 7
Results of Experiment 4: Proportion of targets produced in initial generation/recall phase and proportion of targets recognized (hits) and lures mistakenly identified (false alarms) on the final recognition test.

Condition	Proportion produced	Probability "old"
Generate group		
Read	–	.74 (.03)
Generate	.77 (.02)	.81 (.04)
Lures	–	.28 (.04)
Recall group		
Read	–	.76 (.02)
Recall	.77 (.02)	.90 (.01)
Lures	–	.31 (.03)

Note: Standard errors are in parentheses.

Table 8
Fates of individual items in Experiment 4: Joint probabilities between initial generation/recall and identifying targets as "old" on the final recognition test.

	C_1C_2	C_1N_2	N_1C_2	N_1N_2
Target completions (correct)				
Generate	.68 (.03)	.09 (.03)	.14 (.02)	.10 (.01)
Recall	.74 (.02)	.03 (.01)	.15 (.02)	.07 (.01)
Alternate completions (incorrect)				
Generate	.09 (.01)	.04 (.01)		
Recall	.08 (.01)	.03 (.01)		

Note: Standard errors are in parentheses. C_1 = items successfully generated or recalled in the initial generate/recall phase. N_1 = items that were not generated or recalled in the initial phase. C_2 = target items called "old" (hits) on the final recognition test. N_2 = target items called "new" (misses) on the final recognition test.

Discussion

Experiment 4 used a mixed-list design and showed positive effects of generation and retrieval practice on a final recognition test. The key finding from the experiment was that retrieval practice also produced significantly better recognition performance than generation. The results lend further support to the idea that intentional retrieval in the retrieval practice condition produced greater item-specific processing than incidental retrieval in the generate condition—and consequently retrieval practice enhanced subsequent retention to a greater extent than generation.

General discussion

These four experiments have clearly established that there is an important difference between generating during learning and retrieving during learning and that the difference originates from retrieval mode. The Generate and Recall conditions in these experiments held all test cues constant and differed only in the instructions given to subjects. Intentional retrieval in the Recall condition consistently produced greater retention than incidental retrieval in the Generate conditions.

One challenge we faced in this research was to create conditions that equated initial performance levels but successfully manipulated retrieval mode. Our task met these criteria. We carried out a final analysis across the four experiments with 160 subjects (80 Generate and 80 Recall) to examine any differences in (1) the initial proportion of correct targets produced, (2) initial response times for producing targets, and (3) initial proportion of alternate completions produced. There was virtually no difference in initial proportion correct in the Generate and Recall conditions (.742 vs. .753, $F(1, 158) = 0.39$, n.s.). There was also virtually no difference in response times in the Generate and Recall conditions (4.45 s vs. 4.58 s, $F(1, 158) = 0.53$, n.s.). The difference in the proportion of alternates produced was very small (1.7%) but approached significance (.129 vs. .112, $F(1, 158) = 2.72$, $\eta_p^2 = .02$, $p = .10$). Thus the procedure was successful at holding all test cues constant, manipulating only the instructions given to subjects, and not dramatically altering initial performance levels across the two instruction conditions.

The advantage of retrieval practice over generation was robust. In Experiment 1 there was no generation effect in final free recall but there was a retrieval practice effect. In Experiment 2 there was a generation effect with a mixed-list design but again there was an even greater retrieval practice effect. Experiments 3 and 4 extended this effect of intentional retrieval to final recognition tests. Although the advantage of recalling relative to generating only approached significance in Experiment 3, it is clear that a consistent advantage of intentional retrieval was obtained in all experiments. Importantly, Experiment 3 also showed that generation and retrieval practice both disrupted retention of order information to an equivalent extent (relative to reading the list intact). These findings support the idea that retrieval practice enhanced item-specific processing to a greater extent than generation—enough to overcome disrupted order processing and produce positive effects on free recall (specifically in the pure-list condition in Experiment 1). The conclusion from this research is that when subjects are in an episodic retrieval mode (Tulving, 1983), consciously attempting to reconstruct the past enhances future retention more than generating knowledge under incidental retrieval conditions and this advantage stems from enhanced item-specific processing evoked by intentional retrieval.

There may be educational implications of this research that await future research with educationally relevant materials and tasks. When students implement a learning strategy they often apply it across an entire set of material (cf. Karpicke, 2009; Karpicke, Butler, & Roediger, 2009; Kornell & Son, 2009). This is akin to what happens in pure-list designs where strategies are applied to entire lists. It is difficult to imagine students implementing something similar to mixed-list designs—for instance by reading half the concepts in a text but generating the other half. Interestingly, prior research has not always found that generation produces an advantage relative to reading in educational contexts (deWinstanley & Bjork, 2004; deWinstanley, Bjork, & Bjork, 1996; Metcalfe & Kornell, 2007). Yet generative learning activities, like generating questions or generating self-explanations, are accepted in education (as they should be) based on the idea that generation helps learners actively construct knowledge (Mayer, 2008; Wittrock, 1989). The concept of practicing retrieval or reconstruction has not yet permeated thought in educational circles. One implication of the present research is that a generative strategy might not produce an advantage relative to reading (as in Experiment 1) but a strategy that involves retrieval practice might prove more effective for learning educational materials.

The finding that both generation and retrieval practice enhance individual item processing but disrupt memory for temporal order information raises the question of how generation and retrieval might affect the processing of other types of associative or relational information. For instance, according to multifactor accounts of the generation effect, generation also enhances memory for cue–target associations (e.g., Hirshman & Bjork, 1988; McDaniel et al., 1988; see too Burns, 1990). In the present experiments neither free recall nor recognition tests would allow us to detect effects of cue–target relational process-

ing. Generation and recall might produce different effects on memory for inter-item associative information, but again the present experiments were not aimed at examining this type of relational processing. Although our data suggest that both generation and retrieval disrupt retention of order information, the two activities might produce different results on other measures of relational processing.

From a functional perspective it makes sense that intentional retrieval would produce greater mnemonic effects than incidental retrieval. When subjects are in retrieval mode and intentionally reconstruct the past they must make a spatiotemporal judgment about an item or event—they must reconstruct what occurred in a particular place at a particular time. If a person consciously and intentionally reconstructs knowledge then the probability that they will need to reconstruct that knowledge again in the future is likely high. It seems adaptive to "increment" that knowledge—or something about the process of reconstructing that knowledge—to facilitate future memory performance.

But the question still remains regarding what is actually meant by "enhanced individual item processing". In the case of retrieval practice, it does not seem that the enhancement is due to "elaboration" in the sense of adding additional features to a memory trace. The trace and its features have already been successfully established, sampled and reconstructed, so it does not seem necessary to add more features to elaborate or enrich the trace (cf. Karpicke & Smith, 2009). When subjects engage in retrieval they use the information available in retrieval cues to reconstruct the past. Retrieval essentially involves a discrimination problem of specifying which candidate features are useful for reconstruction. Rather than adding features to memory traces—a mechanism that might underlie elaboration—practicing retrieval may instead specify which features are necessary to solve the reconstruction problem. Therefore retrieval practice may restrict or constrain the set of features treated as candidates when subjects reconstruct knowledge. Perhaps this occurs to a greater extent when subjects intentionally reconstruct the past than when they engage in incidental retrieval in a generation task. These ideas require much further examination, but they may help understand the nature of retrieval practice and how intentional retrieval differs from generation or incidental retrieval.

At this point there are a variety of explanations for the superiority of retrieval practice to generation and these ideas await further research. What we can say now is that there seems to be an important difference between retrieval and generation and—at least in the particular procedure we used—the locus of the effect is enhanced individual item processing. Intentional retrieval disrupts retention of order information, just as generation disrupts it, but the enhanced item-specific processing under intentional retrieval is sufficient to overcome the disrupted order processing and produce positive effects on subsequent retention, specifically in free recall. Practicing intentional retrieval produces greater subsequent retention than generating targets under incidental retrieval instructions.

Acknowledgments

We thank Siara Saliu, Ben Borgmann, Anna Crow, and Kayla Balensiefer for helping collect the data. We also thank James Nairne and Dan Burns for helpful comments.

References

Ballard, P. B. (1913). Oblivescence and reminiscence. *British Journal of Psychology, 1*.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B., et al. (2007). The English lexicon project. *Behavior Research Methods, 39*, 445–459.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 553–563.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*, 201–210.

Brown, W. (1923). To what extent is memory measured by a single recall? *Journal of Experimental Psychology, 6*, 377–382.

Burns, D. J. (1990). The generation effect: A test between single and multifactor theories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 1060–1067.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*, 306–307.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1563–1569.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*, 268–276.

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13*, 826–830.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*, 633–642.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238). Mahwah, NJ: Erlbaum.

Crowder, R. G. (1979). Similarity and order in memory. In G. H. Bower (Ed.). *The psychology of learning and motivation* (Vol. 13, pp. 319–353). San Diego, CA: Academic Press.

deWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: Implications for making a better reader. *Memory & Cognition, 32*, 945–955.

deWinstanley, P. A., Bjork, E. L., & Bjork, R. A. (1996). Generation effects and the lack thereof: The role of transfer-appropriate processing. *Memory, 4*, 31–48.

Gardiner, J. M., & Arthurs, F. S. (1982). Encoding context and the generation effect in multitrial free-recall learning. *Canadian Journal of Psychology, 36*, 527–531.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*.

Geraci, L., & Rajaram, S. (2002). The orthographic distinctiveness effect on direct and indirect tests of memory: Delineating the awareness and processing requirements. *Journal of Memory and Language, 47*, 273–291.

Graf, P., & Schacter, D. L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 501–518.

Greene, R. L., Thapar, A., & Westerman, D. L. (1998). Effects of generation on memory for order. *Journal of Memory and Language, 38*, 255–264.

Hamilton, M., & Rajaram, S. (2001). The concreteness effect in implicit and explicit memory tests. *Journal of Memory and Language, 44*, 96–117.

Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 484–494.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10*, 562–567.

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior, 20*, 497–514.

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language, 32*, 421–445.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649–667.

Jacoby, L. L. (1996). Dissociating automatic and consciously controlled effects of study/test compatibility. *Journal of Memory and Language, 35*, 32–52.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*, 469–486.

Karpicke, J. D., & Smith, M. A. (2009). *Separate mnemonic effects of retrieval practice and elaborative encoding*. Unpublished Manuscript, Purdue University.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17*, 471–479.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57*, 151–162.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*, 966–968.

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31*, 338–368.

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17*, 493–501.

Mandler, G. (1969). Input variables and output strategies in free recall of categorized lists. *American Journal of Psychology, 82*, 531–539.

Mayer, R. E. (2008). *Learning and instruction* (2nd ed.). Upper Saddle River, NJ: Pearson.

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review, 15*, 237–255.

McDaniel, M. A., Riegler, G. L., & Waddill, P. J. (1990). Generation effects in free recall: Further support for a three-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 789–798.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*, 200–206.

McDaniel, M. A., Waddill, P. J., & Einstein, G. O. (1988). A contextual account of the generation effect: A three-factor theory. *Journal of Memory and Language, 27*, 521–536.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review, 14*, 225–229.

Mulligan, N. W. (2002). The generation effect: Dissociating enhanced item memory and disrupted order memory. *Memory & Cognition, 30*, 850–861.

Mulligan, N. W., & Lozito, J. P. (2004). Self-generation and memory. In B. Ross (Ed.). *Psychology of learning and motivation* (Vol. 45, pp. 175–214). San Diego: Elsevier.

Nairne, J. S. (2006). Modeling distinctiveness: Implications for general memory theory. In R. R. Hunt & J. Worthen (Eds.), *Distinctiveness and memory* (pp. 27–46). New York: Oxford University Press.

Nairne, J. S., Riegler, G. L., & Serra, M. (1991). Dissociative effects of generation on item and order retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 702–709.

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 14*, 187–193.

Payne, D. G. (1987). Hypermnesia and reminiscence in recall: A historical and empirical review. *Psychological Bulletin, 101*, 5–27.

Postman, L. (1972). A pragmatic view of organization theory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 3–48). San Diego, CA: Academic Press.

Roediger, H. L., & Blaxton, T. A. (1987). Retrieval modes produce dissociations in memory for surface information. In D. Gorfein & R. R. Hoffman (Eds.), *Memory and cognitive processes: The Ebbinghaus centennial conference* (pp. 349–379). Hillsdale, NJ: Erlbaum.

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210.

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

Roediger, H. L., & McDermott, K. B. (1993). Implicit memory in normal human subjects. In F. Boller & J. Grafman (Eds.). *Handbook of neuropsychology* (Vol. 8, pp. 63–131). Amsterdam: Elsevier.

Roediger, H. L., Weldon, M. S., Stadler, M. L., & Riegler, G. L. (1992). Comparison of two implicit memory tests: Word fragment and word stem completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 1251–1269.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*, 501–518.

Schacter, D. L., Bowers, J., & Booker, J. (1989). Intention, awareness and implicit memory: The retrieval intentionality criterion. In S. Lewandowsky, J. C. Dunn, & K. Kirsner (Eds.), *Implicit memory: Theoretical issues* (pp. 47–65). Hillsdale, NJ: Erlbaum.

Schmidt, S. R., & Cherry, K. (1989). The negative generation effect: Delineation of a phenomenon. *Memory & Cognition, 17*, 359–369.

Serra, M., & Nairne, J. S. (1993). Design controversies and the generation effect: Support for an item-order hypothesis. *Memory & Cognition, 21*, 34–40.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 592–604.

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Langauge, 26*, 589–607.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 210–221.

Tulving, E. (1964). Intratrial and intertrial retention: Notes towards a theory of free recall verbal learning. *Psychological Review, 71*, 219–236.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175–184.

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford University Press.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 359–380.

Underwood, B. J. (1964). Degree of learning and the measurement of forgetting. *Journal of Verbal Learning and Verbal Behavior, 3*.

Underwood, B. J. (1969). Attributes of memory. *Psychological Review, 76*, 559–573.

Wheeler, M. A., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*, 571–580.

Wittrock, M. C. (1974). Learning as a generative activity. *Educational Psychologist, 11*, 87–95.

Wittrock, M. C. (1989). Generative processes of comprehension. *Educational Psychologist, 24*, 345–376.